U.S. Department of the Interior
U.S. Geological Survey

The National Map Catalog Technical Discussion Paper

Quality Assurance of Cartographic Data in Catalog Operations, FY2004

Larry Moore

Mid-Continent Mapping Center
April, 2004

Table of Contents

1 Introduction

The early pilot projects of *The National Map* involved virtually no cartographic review by the USGS. Through FY2003, *The National Map* pilot projects accepted almost any data offered by partners.

This began to change in the later part of FY2003 as population of the Catalog database moved out of a research environment and toward a production environment. By January 2004 both MCMC and ERG had Catalog Support Teams of cartographers and cartographic technicians to populate and maintain the Catalog. These teams are working to formalize procedures for certain types of quality assurance (QA).

The types of QA reported here may strike the reader as imprecise or non-rigorous. For reasons discussed later in this paper, rigorous statistical quality control of *The National Map* data is largely problematic, at least for the near term. But some types of quality inspection and improvement are worthwhile **in spite of** not being statistically rigorous. In traditional topographic mapping, editorial inspection of contour shapes and text placement are examples of this kind of QA. In the distributed digital world of *The National Map* such inspection has different forms and purposes, but qualitative and subjective QA still has a place.

2 Current QA Processes

Following are some types of quality evaluation or enhancement that are becoming standard practice. Most of these processes have been defined only in the last few months, and are still new and immature.

2.1 Inclusion or Exclusion

One of the simplest types of quality improvement is the decision to include or not include a particular layer in *The National Map*. If a layer doesn't contribute new data, adding the layer to *The National Map* only increases clutter in applications. An example is hydro data derived directly from USGS DLGs and served by other organizations.

Another reason for excluding layers is that the data falls outside the scope of *The National Map*. For example, wildlife, mineralogical, and ecological data.

Consistent and sensible decisions tend to promote a product that is easy to interpret. "Bad" decisions about layer inclusion tend to promote clutter, incompleteness, or both.

2.2 Theme and Subtheme Association

The USGS presents data in *The National Map* by theme (hydrography, transportation, elevation…). The premise is that this classification of data will make *The National Map* applications higher quality.

Other organizations do not necessarily organize their data this way, so partner WMS layers must be associated with USGS themes and subthemes. This requires human intelligence and is necessarily manual. Establishing reasonable relationships between WMS layers and *The National Map* themes improves the quality of the overall system.

2.3 Geographic Extent

The National Map intends to build a national application from a "quilt" of data sources. Many of these sources are not national in scope, but cover only one state, one county, or one city. To behave reasonably, applications must know the geographic extent of every data layer. Without knowledge of extents, an application has little choice but to make all layers available to the user at all times. In a viewer application, this results in (for example) layers for California being offered as choices even though the user is zoomed in on New Jersey. *The National Map* contains about 800 layers already[1]. Having them all available for display at all times would make a viewer application unusable.

Data providers are usually concerned with exactly one geographic extent (their State, county, or whatever), so their services and applications can ignore the issue. Adding and managing footprints is a necessary quality step for *The National Map*.

2.4 Scale

Digital applications can display data over an extremely wide range of scales. *The National Map* currently contains data with nominal map scales as small as 1:3,500,000 and as large as 1:5,000. Maintaining layer-specific information about data scale improves overall quality by giving applications guidelines for when it is appropriate to allow a layer to be displayed.

2.5 Metadata

The USGS has for years strongly endorsed the FGDC metadata standard. Insisting that all layers of *The National Map* have appropriate metadata is a substantial contribution to quality. The program provides education, training, and tools to partners to help them create metadata for *The National Map* data.

The catalog support teams review the metadata of every WMS layer harvested into *The National Map*. This review is not exhaustive; it is intended to insure that metadata files are coherent and free of blunders (for example, that a metadata file actually describes the layer it is associated with).

A firm policy on metadata was set only in January, 2004. At this writing, about 40% of layers in *The National Map* still lack metadata links. The current goal is to achieve 100% compliance by May 15.

A similar issue, though one specific to *The National Map* viewer application, is the presence of legends. The catalog support team also works with data providers to create consistent legends.

2.6 Availability

The National Map depends on hundreds of data layers being served from more than 80 services[2], most of which are not owned or controlled by the USGS. When one of these services is not available (due to software or hardware failures, system maintenance, or other causes), the data obviously cannot be seen in *The National Map*.

[1] 776 public layers as of April 5, 2004
[2] 86 services as of April 5, 2004

Service availability is continuously monitored by USGS software. During normal business hours, the output of these programs is monitored by the catalog support teams. When a service has been unavailable for an hour, we contact the owner of the service. It is often the case, especially for local GIS organizations, that the owner is not aware their service is down.

This monitoring and notification is an early stage of a much bigger and longer-term effort to improve the overall reliability of *The National Map* through stronger policies, automated monitoring, and automated service failover.

The MCMC catalog support team has published two reports summarizing service availabilities in FY2004. These analyses are a type of statistical quality control in *The National Map*.

2.7 Layer order and transparency

Vector layers must have transparent backgrounds to be useful for an application such as *The National Map* viewer. Transparency must be provided by the source service, but it is monitored by the catalog support team.

Correct "stacking" of layers is sometimes important, especially for raster data. There are instances in *The National Map* of (for example) State image data that is more recent than the national coverage, and county data that is more recent than the State coverage. In such an instance, it is important to display the smaller footprints on top of the larger ones. The catalog service calculates and delivers this information to assist applications in proper ordering. The catalog database contains provisions to allow the calculated ordering to be overridden for specific layer combinations when the need arises.

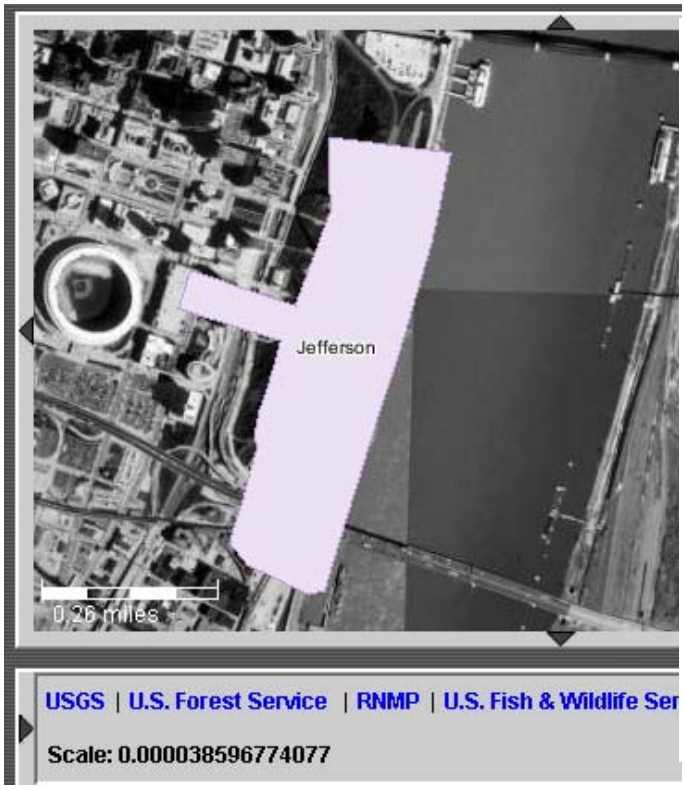
2.8 Horizontal registration with other layers

"Horizontal registration" is a weak form of horizontal accuracy (stronger forms are discussed below). The type of registration inspection currently performed does not make any statistical statements about absolute horizontal position. Rather, it assumes that most data providers supply reasonably accurate data to *The National Map*. When data layers from independent sources are overlaid, the resulting display should be free of glaring registration discrepancies. Most of the time this is in fact true, and serious errors in horizontal position therefore tend to stand out.

Careful inspection of images is not quite as simple as it sounds. *The National Map* already has some 800 layers, most of which have nominal map scales of 1:100,000 or larger and cover areas at least as big as a State. A complete visual inspection of any one layer against all the other layers with overlapping viewscales and extents is not a small task.

On the other hand, *The National Map* is an excellent environment for performing this type of quality assurance, precisely because it includes so many independent data layers. The catalog support team began doing systematic inspection of data only in January 2004. Several serious positioning errors have been found since then, most of them previously unknown to the partners that own the data.

This type of inspection typically finds a lot of "low hanging fruit." The benefit/cost ratio of the first hour of inspection is quite high. Deciding when to stop is a harder problem. There is also the separate, mostly political, problem of what to do about errors when you find them; data owners do



The solid polygon is the Jefferson National Park (Gateway Arch) in St. Louis, MO. It is in a layer provided by the National Park Service (NPS) showing all National Park extents. The park is about 200 meters south of its correct position relative to the DOQ image (shown) and all other layers. Measurements in other parts of the country demonstrated the NPS error is haphazard. In some areas it is as large as 2,000 meters, in other areas almost negligible. There is no pattern to either the direction or the magnitude. The Park Service has been notified. While waiting for their response, the problem has been "fixed" by changing the minviewscale value so the layer will only display at scales much smaller than this display. In spite of the horizontal inaccuracy, the layer still has value as a small-scale reference.

For reasons explained in the text, this is not true statistical process control. Nevertheless, the benefits of this type of inspection and layer comparison are fairly obvious.

not necessarily agree that a particular error is significant, or may have many good reasons why the data cannot be corrected immediately.

2.9 Content consistency between layers

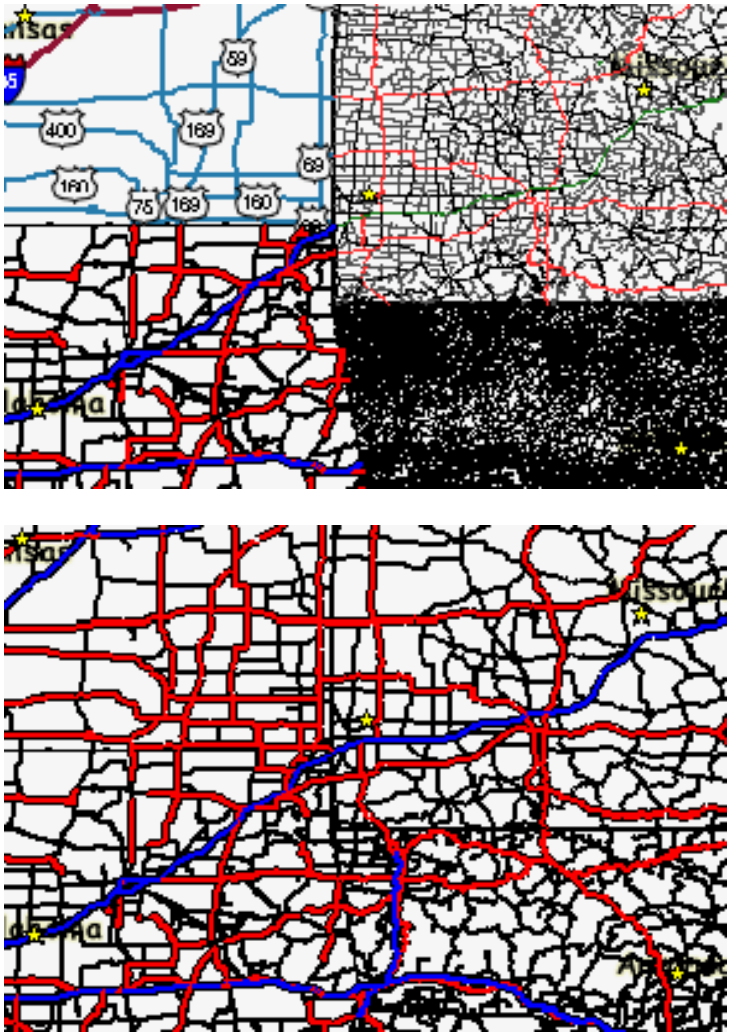
Just as registration inspection is a weak form of horizontal accuracy testing, cartographic consistency inspection is a weak form of content verification. For example, geographic names in the GNIS layer should tend to line up with features in the elevation, hydrography, boundary, or other layers. Boundaries should tend to match coordinate grids or the shapes of features in other layers. Man-made features tend to form recognizable patterns.

It is difficult to define standards or algorithms for this type of quality assurance, but human intuition and experience make it useful anyway. As with registration inspection, the benefit/cost ratio is high at first, but drops rapidly with time.

2.10 Standard symbols

This is not an issue of inspection or quality evaluation. Like theme association or metadata creation, standardizing symbols actually adds quality to *The National Map*.

The mechanism for symbol standardization is Styled Layer Descriptors (SLD). SLDs are an extension of the OGC WMS specification. They allow an application to override the symbols provided by the data service, and apply customized symbols based on feature attributes. This requires a conceptual mapping between the source data attributes and the desired standard symbol



The top image displays the road symbols provided by the State services at the corner of KS-OK-AR-MO. The bottom image is the same area with styled layer descriptors (SLD) applied.

Using SLDs degrades the original data by collapsing the States' attribute models to a kind of lowest common denominator. Nevertheless, the enhancement to visual clarity is so great that the application of SLDs must be regarded as a quality improvement. One way to look at this is that the **signal** of road information is reduced slightly by collapsing the attribute models, but the visual **noise** is reduced a huge amount by standardizing the symbols. The signal-to-noise ratio is therefore much more favorable in the bottom image.

The SLDs are applied only in *The National Map* viewer. The original data are not disturbed, nor are any other applications that use either the default symbols or different SLDs.

set. The mapping is necessarily manual, but need only be done once for each data source (provided the source attribute model and the target symbol set are both stable).

This is a very powerful tool. It contributes to constructing a national "blanket" from the "quilt" of State and local WMSs. Except for the one-time attribute mappings, the blanket construction is automatic. SLDs do not disturb the original data or any existing applications that use different symbols. Showing consistent symbols across political boundaries creates the illusion of a true national map.

A significant SLD pilot project was completed in February, involving three States (MO, KS, AR) and two layers (roads, boundaries).

3 Prospects for True Statistical Process Control

The QA steps taken so far are primarily qualitative, ad-hoc, and based on visual inspection of data. An important question is how these procedures could be enhanced to achieve true statistical process control.

Statistical process control can, at least in principle, be applied to any cartographic process. However, in the case of *The National Map*, formal statistical control would likely be extremely expensive. The basic reason is fairly simple: statistical process control requires that measurements be made against a control dataset. The control must be both independent of, and at least as accurate as, the thing being measured. For example, estimating the positional accuracy of a new digital orthophoto usually requires collecting new GPS points in the field. This is because the new orthophoto was probably created using the best horizontal control available. Testing against the same control tells us nothing, because that control is not independent of the product. Testing against older control is not adequate because older control is probably less accurate than the product.

3.1 Horizontal Accuracy Testing

Collecting new control for testing *The National Map* data is probably out of the question. Field operations to collect new control are extremely expensive, and the rapid rates of growth and change in *The National Map* almost certainly preclude any independent control program by the USGS or anyone else.

Nevertheless, a limited type of horizontal accuracy testing could be implemented using the USGS DOQ layer as control. This is still a weak form of accuracy testing, for at least three reasons:

1. It works only if there is agreement that the DOQ layer is sufficiently accurate.
2. It works only for data layers that contain features visible on aerial photos.
3. Because point comparisons must be between features visible on photos, true randomization is difficult and probably very expensive. This means the statistical results will always be open to challenge.

In many ways, this would be just an enhanced and more expensive version of the qualitative inspection we are doing now. It might be worthwhile anyway, because making measurements and doing statistical analyses makes the process explainable and the results repeatable. It also promotes a thoroughness that is hard to achieve from ad-hoc, qualitative inspection.

3.2 Attribute Verification

Testing the quality of cartographic attributes is even harder. Unlike horizontal position, it is generally not possible to use images to check the accuracy of attributes. You cannot, for example, find route numbers or geographic names on an aerial photo.

The National Map aspires to use the best available data, everywhere, all the time. To test the accuracy of attributes you need an independent dataset, with comparable content, that has higher accuracy. But if such a dataset is known to exist, it becomes the best available data, the original dataset becomes worthless, and the testing problem remains.

It is hard to imagine any USGS-owned process that would solve this problem in a formal, statistically defensible way, while not costing a fortune in time and money. The USGS can do some checking for consistency between layers, but responsibility for absolute accuracy of attribute data must rest with the original data producer.

4 Conclusions

The National Map is a data integration and data delivery system, not a data production system. Integration has different quality assurance objectives than primary data production, which we are beginning to understand and address.

It is hard to see how *The National Map* can ever guarantee the absolute accuracy of either positions or attributes of the cartographic data it uses. For the most part, assurances of completeness and accuracy must be given by the organization that produced the GIS layer from primary sources. Producers can be USGS national data programs, State or local government agencies, or private companies.

The National Map can do some quality assurance based on comparisons of layers to each other, as opposed to independent and superior data sources. *The National Map* can also add value by notifying providers of inconsistencies between sources, applying standard symbol sets, and adjusting scales and other display attributes as appropriate for a national application. Such processes are being studied and implemented now, but will probably never be finished or stable. Adjustment and improvement of the data inspection processes will always be important.